LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-655379

Initial Global Seismic Cross-Correlation Results: Implications for Empirical Signal Detectors

D. A. Dodge, W. R. Walter

June 5, 2014

Bulletin of the Seismological Society of America

Disclaimer

Initial Global Seismic Cross-Correlation Results: Implications for Empirical Signal Detectors

D. A. Dodge[1], W. R. Walter[1]

[1]Lawrence Livermore National Laboratory
7000 East Ave, Livermore, CA 94550, USA

May, 2014

Douglas Dodge: dodge1@llnl.gov
William Walter: walter5@llnl.gov

Corresponding author:
Douglas Dodge
Lawrence Livermore National Laboratory
7000 East Avenue
Livermore, CA 94550
MS 046
925-423-4951

Abstract

In this work we cross-correlated waveforms in a global dataset consisting of over 310 million waveforms recorded between 1970 and 2013 for two purposes: to better understand the nature of global seismicity and to evaluate correlation as a technique for automated event processing. We found that about 14.5% of the events for which we have at least one waveform correlated with at least one other event at the 0.6 or higher level. Within the geographic regions where our waveform holdings are complete or nearly complete, that fraction rose to nearly 18%. Moreover, among the events for which we had one or more seismograms recorded at distances less than 12 degrees, the fraction of correlated events was much higher, often exceeding 50%.

These results imply that global seismicity contains a large number of "repeating" events, that is, events which are sufficiently similar to each other to have correlated waveforms over the time period spanned by our dataset. These results also are very encouraging for using correlation in aspects of automated event processing. It is well known that because of the strongly implied similarity of the sources of correlated signals, they can be used as empirical signal detectors (ESD), to detect, locate and identify an event using as few as one channel. The results reported here are very encouraging for using correlation and perhaps other forms of ESD for regional network processing and continental global processing since, for example, nearly all continental seismicity is within 12 degrees of at least one International Monitoring System station.

Introduction

It has long been known that seismic events can produce seismograms with strong similarity to previously recorded events. Quantitatively this characteristic of seismicity is often measured through waveform correlation. High correlation values between seismograms from different events imply these events have similar locations, mechanisms and other properties. Strong seismogram correlation, when it occurs, can thus be extremely useful in seismic event processing, as well as shedding light on seismic properties such as slip recurrence rates on fault patches. In this paper we attempt to better quantify how much of the Earth's seismicity is correlated and how such correlation is distributed in space and time.

Since at least the 1960's it has been known that correlation can be used as the basis for highly sensitive detectors (e.g. Anstey, 1966; Van Trees, 1968). The literature has many examples of correlation detectors applied to tightly clustered seismicity observed at local to near-regional distances; e.g. (Israelsson, 1990, Harris, 1991, Gibbons and Ringdal, 2004, 2005) to name a few. Using array-based correlation detectors, Gibbons and Ringdal (2006) demonstrated an order of magnitude reduction in the detection threshold relative to incoherent detection on a beam. These uses of correlation are so well established that at the U.S. National Data Center (USNDC), correlation detectors are routinely used for repeating sources (Junek et al., 2013). Here we treat correlation as one type of Empirical Signal Detector (ESD), a term coined by Junek et al., 2013 to refer collectively to pattern matching detectors such as correlators, subspace detectors (e.g. Harris, 2006), and matched field detectors (e.g. Harris and Kvaerna, 2010).

Correlation detectors have also been applied with some success to earthquake aftershock sequences. Large earthquake sequences are a problem for monitoring agencies because the high rate of activity can make it difficult for analysts to keep up with processing deadlines. This is due to the shear volume of events to be processed and to the numerous false associations produced by current automated systems under conditions of high seismicity. If it is common for a significant fraction of events to be correlated, then a pipeline suitably designed to use correlators to pre-group detections and prevent many false associations could far out-perform current systems especially during large aftershock sequences.

Harris and Dodge (2011) have used correlation in combination with subspace detectors in an automated system to track events in an aftershock sequence. They demonstrated a potential analyst workload reduction of up to 73%. Slinkard et al., (2013) applied correlation detectors to three aftershock sequences using stations from 27 to 900 km distant. They found that the percentage of bulletin events detected by correlators ranged from 30% to 92%. These examples are encouraging, but not definitive.

Correlation detectors have also been shown to be effective over much larger regions. For example Schaff and Richards (2004, 2011) discovered that about 13% of 18,000 earthquakes recorded at regional distances in China were sufficiently well correlated that they could be detected and located using waveform correlation. We

find that the global average is about 18% and at short distances can rise to 50% or more. Furthermore, there is potential for higher-rank subspace detectors to improve considerably on the detection rates of pure correlators. Automated processing of 18% of world seismicity would be a significant reduction in analyst workload and the percentage of events detected by ESD is expected to grow over time. Also, a suitably designed system could mask or cancel the signals associated with all its detections. This could considerably ease the workload on the associator at times of high seismicity, resulting in fewer false associations. For these reasons it seems worthwhile to consider the use of correlators or more advanced empirical signal detectors as part of future global pipeline systems.

The present computational costs appear to be high, relative to current practice in seismology, but not by the standards of "Big Data" practitioners. For example, all channels of the IMS seismic sensors produce only a few tens of gigabytes of data per day. By comparison, in 2013 the Facebook data warehouse took in 500 terabytes per day (Miners, 2013). Implementing a system on that scale would be expensive today. However, the strong competition among vendors virtually assures that a system designed in a few years will be able to take advantage of commodity solutions with more than enough storage and processing power.

In a future paper we will examine some of the hardware and software issues involved in scaling correlation detection to an operational capability in a global pipeline. In this paper we describe how effective is expected to be; e.g. can we

125 better quantify how much of the Earth's seismicity is correlated and how it is
126 distributed in space, time and with what event characteristics. In this paper we
127 attempt to answer these questions by cross correlating a large, globally distributed
128 set of seismograms and analyzing the statistics of the resulting set of correlations.
129

130 ## The Dataset

131 Lawrence Livermore National Laboratory (LLNL) operates a database of seismic
132 events and waveforms for research on nuclear explosion monitoring and other
133 applications. The waveforms are digital time series of ground motion recorded by
134 seismometers installed at seismic stations. Typically, the seismometers produce
135 output on multiple channels corresponding to different orientations and pass bands,
136 so that often the same events are recorded on multiple channels at each station.
137

138 The LLNL database contains nearly 3.8 million events associated with more than
139 310 million waveforms at nearly 6,300 stations (Figure 1). The events are compiled
140 into a reconciled list from tens of individual bulletins produced by seismological
141 organizations around the world (e.g. USGS, CTBTO, ISC, numerous regional and local
142 network operators). The waveforms come from the same sources and especially
143 data collection centers such as the Incorporated Research Institutions in Seismology
144 (IRIS) Data Management Center (DMC). The figure (A) shows the completeness of
145 waveform holdings geographically. The figure was produced by gridding the Earth's
146 surface into 50km by 50km cells and, within each cell, dividing the number of events
147 for which we have at least one seismogram by the total number of events in our

148 catalog for that cell. The color scale indicates the completeness; with black
149 indicating no waveforms and white indicating that for every event in the cell we
150 have at least one waveform. Although the data set has global coverage, the
151 completeness is highest in the Middle East, Eurasia, Fennoscandia, and Western
152 North America. Many of the conclusions reached in this work are based on analysis
153 of data from the regions where our coverage is 80% or greater. By restricting our
154 analysis to this subset of the data we hope to minimize biases resulting from uneven
155 distribution of waveforms in the database. The waveforms in the LLNL database
156 span a period of time greater than 60 years (B), but the earliest data are for stations
157 and channels not found later. In fact, the effective time period for correlation
158 processing is about from 1970 to the present (C).
159

160 ## Procedure

161 In order to investigate the correlation behavior of seismic signals over a wide range
162 of seismic wave types and frequencies we correlated catalog events in 8 seismic
163 phase windows (e.g. P, S), as well as in 15 frequency bands for each window. The
164 bands and windows used are detailed in Tables 1 and 2. Correlations are performed
165 for data recorded on a common station and channel (STA-CHAN hereafter). It is
166 impractical and unnecessary to calculate correlations for all possible event pairings
167 per STA-CHAN. For our data set this would have required the calculation of over
168 10^{15} cross correlations. Rather, it is sufficient to calculate cross correlations only for
169 those event pairs that we know to be close enough spatially that they might produce
170 correlated seismograms. From preliminary studies we determined that it was rare

for two events with correlated seismograms to have relative mislocations of more than 50 km so we chose that distance as a search radius. Although restricting the calculation of correlations only to nearby events dramatically reduces the number of correlations which must be calculated, with 3.8 million events to compare it is very important to have an efficient strategy for finding nearest neighbors. We employed a Java Spatial Index, which is the Source Forge implementation of an R-Tree (Guttman, 1984). For each STA-CHAN we retrieve all events recorded by that STA-CHAN, and use the R-Tree to build 'islands' of events within 50 km of one-another and process all pair-wise combinations in the island.

Processing of an island is shown schematically in Figure 2. An arbitrary event is chosen as the starting point and the R-Tree is used to find all neighbors within 50 km. After measuring correlations with those neighbors, the event is removed from this list and the processing is repeated with one of its neighbors. Eventually an event with no neighbors is found, and the island is completely processed. The processing of an event pair within an island is shown schematically in Figure 3. The waveforms are retrieved (as required) and the possible windows and bands are identified. For each phase and band, the seismograms are filtered and trimmed, and a signal-to-noise ratio (SNR) test is performed on each window. If both windows pass the test, they are correlated and if the correlation meets or exceeds 0.6, the results are written to the database correlated event list.

In all over 650 million correlations were written in about 42 days on a configuration consisting of 4 servers with 44 cores and 613 gigabytes of RAM. In addition to the correlations that were written to the database, about 700 million correlations were computed but rejected. SNR tests removed nearly 135 million windows from processing before a correlation was computed. There were nearly 678 million cases where a band was skipped because the sample rate was too low or the window was too short for the band (i.e., the window failed a simple test to prevent low time-bandwidth-product correlations). Subsequently, we re-implemented the correlation processing code using Hadoop (an open-source framework for processing large-scale data sets using commodity clusters) and achieved a speedup of nearly a factor of 20 on a test subset of events. The Hadoop implementation will enable larger and more complete investigations into correlation behavior in the future. Details of the faster Hadoop implementation are described in detail in the Addair et al. (2014) paper.

We performed post-processing to remove correlations due to signal artifacts. A significant number of seismograms used in this study contained artifacts that correlate quite well. The data from some stations was so contaminated, that tens of millions of correlations were due to artifacts. Examples are shown in Figure 4. To identify and remove segments with these artifacts we used a random forest classifier (Breiman, 2001). Classification was a 2-step process; operating first on a set of 16 raw waveform features and then on a set of 8 filtered features. The classifier was trained using a data set of 18,300 randomly selected and filtered

windows, which were manually reviewed and classified. Based on 10-fold cross validation testing, the classifier achieved about 95% precision in classification. After classification 371,209,733 correlations were retained.

General Characteristics of the Correlation Results

In all, 14.5% (542,405) of the 3,745,879 distinct events in our waveform table had valid correlations that met or exceeded the 0.6 acceptance threshold. Nearly 40% of the 6,266 stations produced at least one valid correlation. Figure 5 shows the distribution of the retained correlations by phase (A) and by band (B). Most of the correlations are for the whole waveform and for the S phase. Between them they account for nearly 271 million (~73%) of the correlations.

The whole-waveform window started 10 seconds before the theoretical P-wave arrival and continued to MIN (Δ_{km} / (3 km/sec), 2000 sec). Because most of the retained correlations were for relatively short event-station separations, the average length of the whole-waveform window was about 82 seconds. The effectiveness of the whole-waveform window relative to shorter windows designed to extract single phases is somewhat surprising. We initially suspected that the correlation classifier had disproportionately removed shorter windows based on time bandwidth product values. However, examination of the removed correlations showed that the whole-waveform window was most often removed, followed by the Sn and S windows. A more likely explanation for the predominance of this window in our results is that it always exists, whereas the other windows only are computed

239 if they are predicted by the AK135 travel time calculator for the event-station pair.
240 Furthermore, the whole-waveform window always samples the part of the
241 seismogram with the highest SNR whereas specific phase windows often do not.
242
243 The correlation results also are predominantly short period. Figure 5(B) shows the
244 number of correlations as a function of filter band. The 1-2 Hz band is by far the
245 most productive band. Most of the remaining correlations are in bands centered
246 around or above 1 Hz. The majority of correlations were for signals recorded at local
247 to regional distances, and at these distance ranges, (and also for teleseismic P) these
248 are the filters one would expect to be most effective at bringing out the desired
249 signal. Because we did not compute correlations for windows containing fewer than
250 10 cycles of a signal at the dominant period in any given band, there are no
251 correlations in long-period bands at local distances or for any window other than
252 whole-waveform. This could also contribute to most correlations being for the
253 whole-waveform window.
254
255 Figure 6 shows the correlation counts as a function of event-station separation for
256 long period bands (A), mid period bands (B) and short period bands (C). The
257 correlations in (A) are primarily of surface waves recorded in long windows, so
258 except for the band (0.5 – 1.0Hz) there are no observations at very short distances.
259 This is a side effect of our windowing strategy as discussed previously. At mid to
260 short periods, the dominant feature in the plots is a drop in numbers of correlations
261 of about 3 orders of magnitude for distances greater than 8 to 10 degrees. From

that point to about 90 degrees, the number of correlations stays relatively constant except for a bump between 35 and 51 degrees.

This behavior was surprising since our expectation was that with increasing frequency, attenuation of the signal would cause decreasing correlation values with distance. To be sure that the correlations seen at teleseismic distances were not dominated by misclassified artifacts we performed a manual inspection of a subset of the teleseismic results. Examination of 100 seismogram pairs chosen randomly from the correlation results for distances of 30 to 90 degrees in the mid period and short period bands showed that in all bands except one, every sample contained valid seismograms. Interestingly, nearly all these teleseismic data are recorded by IMS arrays. The increase in the correlation counts between about 35 and 51 degrees is a real feature. It turns out that a handful of arrays are situated such that several major seismic zones fall within that distance range for these arrays. This is indicated in part (D) of Figure 6.

Figure 7 shows the magnitude differences (left) and the distribution of time separations (right) for correlated event pairs in our results. The data are divided into four bins based on the average magnitude of each event pair. Panel (A) shows results for Mw <= 2. Panel (B) shows results for 2 < Mw <= 4. Panel(C) shows results for 4 < Mw <= 6, and panel (D) shows results for 6 < Mw <= 8. The data were prepared by selecting all event pairs in the correlation results table for which the whole-waveform correlation exceeded 0.6 in one or more high-frequency (>0.5 Hz)

bands. We are interested in understanding the detection characteristics of whole-waveform, high-frequency templates, and by down-sampling the data we hope that the resulting statistics will be more representative of that population. The repeat interval plots were produced using these data.

Our first attempt at producing the magnitude difference distributions yielded histograms with surprisingly heavy tails. Examination of the outliers revealed that in nearly all cases, one or both of the events being compared had only a single magnitude estimate from a local or regional bulletin, and a very large number of these appear to be off by a magnitude unit or more. Accordingly, we decided to remove all event pairs for which the only magnitude estimates are from single local or regional bulletins. This significantly reduced the number of event pairs, but there are still thousands in each magnitude range. The resulting magnitude difference histograms show that over the entire span of magnitudes in our database, events are likely to correlate well at short periods only if their magnitudes differ by less than two units.

The histograms of repeat intervals were produced by binning the time differences of correlated events in the 4 different magnitude ranges. The most obvious feature of these plots is the abrupt ending just short of short of 20 years. This seems surprising since the time span of the waveform data is about 40 years. However, as Figure 8 shows, the LLNL waveform data can really be thought of as two distinct sets that share only a few tens of STA-CHAN between the epochs of (1970 – 1990) and (1990

– Present). At larger magnitudes, the repeat frequency decays with interval length as it must, but for Mw < 4 there is a flattening of the slope starting around 7 or 8 years. This appears to be an artifact of the way we have built our research database over many years: initially disk space limits caused us to use a short distance threshold for M<4 data collection, whereas more recently we have been collecting globally without magnitude or distance thresholds. For the largest magnitude event pairs (D) there is about an order of magnitude increase in the number of repeats in the shortest-duration bin. These are almost entirely aftershocks recorded at teleseismic distances, correlated using long windows in the 1-2 Hz.

Prevalence and Geographic Distribution of Correlated Events

The geographic distribution of correlation results as fractions of total seismicity is shown in Figure 9. To produce these plots we gridded the Earth's surface into 50km by 50km cells, and in each cell computed the ratio of correlated events to the total number of events reported in bulletins for the time period in which we have waveforms for the cell. Because we are interested in understanding the prevalence and distribution of correlated seismicity, and because the LLNL research database waveform holdings are not complete globally, we restrict most of our analysis to the region outlined by the white dashed lines. Within this region, we have waveforms for nearly all events, and therefore believe that the patterns we see in these regions are not biased by variations in data completeness.

Panel (A) shows the distribution of correlated seismicity without any restriction by band, phase, or magnitude. Globally, all or nearly all of the major seismogenic zones of the Earth are evident. The most striking features within our analysis region are the bright spots in Fennoscandia, central Asia, the Andaman Sea, and Iran. By contrast, the Mediterranean region shows a much lower fraction of correlated events. Some of these regions (e.g. Fennoscandia) have a large amount of mine seismicity which is known to correlate quite well (e.g. Tarvainen and Husebye, 1993). Panel (B) shows the distribution of correlated seismicity for events of magnitude 5 and greater. Within the analysis region, the fraction of correlated seismicity appears to be much larger on average than the distribution in (A) with most areas having a fraction greater than about 0.4. Evidently, the bright spots seen in (A) correspond to areas that have both a high density of low magnitude events and one or more stations close enough to have high SNR recordings for those events. This interpretation is supported by panel (C), which shows the fraction of events for which we have waveforms from stations within 5 degrees of the epicenters. Most of the bright spots in (A) correspond to bright spots in (C), and the Mediterranean is seen to be a region with a relatively low density of nearby stations (in our waveform database).

Evidently, correlated seismicity is not restricted geographically. But are enough events correlated to warrant making correlation detection part of routine pipeline processing? For the entire data set, about 14.5% of the events for which we have one or more waveforms have mutual correlations. Within the analysis region where

353 our waveform coverage is mostly complete, the fraction increases to nearly 18%
354 and the ratio of correlated events to events reported in bulletins is nearly as high
355 (17%). Figure 10 shows the fraction of correlated seismicity as a function of source-
356 station separation in different magnitude ranges. The intent is to show how well
357 correlation detectors might perform in a system where the nearest station may be
358 several degrees from the source.
359
360 Panel (A) shows the behavior when using all possible bands and phases. For events
361 with M > 5, an astonishingly large fraction (~0.3 - 0.8) of events are correlated even
362 at very large distances. Many of these are long-period surface wave correlations, and
363 while they do not indicate the events are in close proximity, when detected at
364 multiple stations the correlated arrivals can be used to perform very accurate
365 relative locations (e.g. Cleveland and Ammon, 2013) and this could be used in
366 pipeline processing. Events with M <= 4 only have significant correlation fractions at
367 distances < ~10 degrees. However, for events in the range 4 < M <= 5 and out to
368 about 30 degrees, the correlation fraction varies from 10% to 20%. . About 10% can
369 be correlated to 70 degrees.
370
371 Panel (B) shows the behavior using only short-period bands. The correlation
372 fraction for large magnitude events averages 0.2 to 0.3 over a very large distance
373 range. This is encouraging, but should be interpreted cautiously. Nearly all these
374 correlations are for P in bands 1-2, 1-3, 2-4, and 1-5. Often these signals contain a
375 relatively short P-pulse followed by low-amplitude coda. For example, Figure 11

shows 80s long seismograms recorded at station KK01 for a group of 15 events correlated in the 1-2 Hz band. The correlation windows used at KK01 were about 35s long. Most of the similarity occurs within about the first 20 s. In such narrow-band, short-window cases the correlation can provide excellent relative timing between these P phases but is unlikely to indicate the causative signals are very closely located to each other. More likely, they are separated by a few tens of km. (The bulletin locations indicate a maximum separation of about 70 km.) This level of resolution may still be useful for association, or relative location based on network results but is insufficient for assignment of location based on single-station correlation, for example. Over the remaining magnitude ranges in Figure 10 (B) the behavior is similar to that of (A): The correlation fraction is large at less than ten degrees, and only the magnitude 4-5 events have a significant correlation fraction at greater distances.

Panel (C) shows the behavior in short-period bands and using a correlation threshold of 0.8. Such conditions might be required if the correlations are to be used to offload work from the associator by directly classifying new events. With these restrictions, a significant fraction of events that correlate can only be found at distances less than about 8 degrees.

Utility of Correlation Detectors for Global Seismic Monitoring

Clearly, correlation detectors (and ESDs in general) can be expected to be useful for local to regional monitoring systems. This is, after all, the domain in which many

399 successes have been reported, and is the distance range in which this study finds the
400 greatest fraction of correlated waveforms. In addition, our results suggest that ESDs
401 can play an important role in a global monitoring system as well. For example, the
402 International Monitoring System (IMS) of the Comprehensive Nuclear-Test-Ban
403 Treaty Organization (CTBTO) will have, when complete, 50 primary and 120
404 auxiliary seismic monitoring stations (Brely, 2010). The station density of the IMS is
405 such that a large fraction of the Earth's continental seismicity is within 12 degrees of
406 at least one IMS station. This is shown in Figure 12. In the figure, the small circles
407 are each centered on an IMS station and have radii of 12 degrees. Each panel shows
408 the Earth's seismicity color-coded according to distance from the nearest station.
409 The top shows the situation when just the primary stations are used and the bottom
410 shows the situation using both primary and auxiliary stations. When all stations are
411 considered, a very large fraction of the Earth's seismicity is found to be within 12
412 degrees. Of course, this does not take into account ambient noise levels and other
413 factors that may make a station less useful. But it does suggest that a very large
414 fraction of the IMS stations may perform usefully in an ESD subsystem. The actual
415 design of a full-scale ESD subsystem for a large network such as the IMS would be a
416 complex undertaking and is beyond the scope of this paper.
417
418 **Discussion**
419 In order to understand better the characteristics of global seismicity and evaluate
420 the utility of seismic waveform correlation in automated event processing systems,
421 we performed a very large scale global cross-correlation on a research database

containing more than 300 million seismic waveforms. To understand better the dependence of waveform correlation behavior on time-bandwidth characteristics we performed the correlations in multiple time windows and frequency bands. After eliminating problematic non-seismic signal waveforms, we created a database table with about 371 million correlated seismograms. We are still examining these results in detail. In this paper, we described the most general characteristics of the results: the time, frequency, distance, and magnitude relationships between the events that showed strong correlation. In particular we are motivated by the potential to use such waveform correlation characteristics in future automated processing systems, both to lower detection thresholds and reduce the workload of human analysts.

A major potential application of seismic waveform correlation would be as part of empirical signal detectors (ESD) (e.g. correlation, subspace, matched-field, etc.). These are well known to be highly sensitive relative to power detectors. In addition, because seismic sources only produce correlated signals if the sources are very similar in location and mechanism, ESDs can detect, locate, and identify a source using as little as one channel. Because of these advantages, ESDs have been considered as components in pipeline architectures. To date, however, there have been no large-scale deployments. The barrier to deployment is high and includes the following factors:

1. Existing pipeline architectures are very mature, and for the most part do their job very well without resort to correlation detection. Operators of these

systems necessarily must be conservative about making major changes to these systems.

2. Although correlation detectors have been shown to work well in a number of regions, heretofore, it is unknown how effective they would be on a global scale.
3. Large-scale correlation processing is computationally expensive, and cannot work on the architectures currently used by pipeline operators.

We did not address the first item, but here point out that the current monitoring architecture is decades old and will eventually need to be replaced. We suggest that any redesign of a pipeline processing system should keep ESD in mind.

This paper primarily focused on the second question, global effectiveness. We found that about 14.5% of the events share at least one waveform correlation with another event (correlation coefficient >= 0.6). Within the geographic regions where our waveform holdings are complete or nearly complete, that fraction rose to nearly 18%. Moreover, among the events for which we had one or more seismograms recorded at distances less than 12 degrees, the fraction of correlated events was much higher, often exceeding 50%. We find these results to be very encouraging, with respect to point 2, since nearly all continental seismicity is within 12 degrees of at least one IMS station.

Finally on the third point on computational expense, the landscape is changing very rapidly. During the course of this work, we became very aware of the computational complexity issues, and particularly of the impact of I/O on processing time. We ultimately re-implemented our correlation processor on the open-source Hadoop

468 platform and found a nearly 20X speed improvement (Addair et al., 2014). The big-
469 data analytics ecosystem of which Hadoop is a part is evolving rapidly and many
470 businesses are processing huge amounts of data in real time using these
471 technologies. We think this will lead to a viable architecture for processing
472 streaming seismic data using correlation in the next few years.
473

474 Acknowledgements

475 We thank Stan Ruppert and Terri Hauk for their long-term work to build and
476 maintain the LLNL research database. We thank Travis Addair for work on the
477 massive correlation processing. We thank Steve Myers and Dave Harris for
478 comments that improved the manuscript. This work was performed under the
479 auspices of the U.S. Department of Energy by Lawrence Livermore National
480 Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National
481 Security, LLC. This is LLNL Contribution LLNL-JRNL-XXXXXX.
482

483 References

484 Addair, T. G., D.A. Dodge, W.R. Walter, S.D. Ruppert, Large-scale seismic signal
485 analysis with Hadoop, Computers & Geosciences, Volume 66, May 2014, Pages 145-
486 154, ISSN 0098-3004, http://dx.doi.org/10.1016/j.cageo.2014.01.014.
487
488 Anstey, N.A., 1966. Correlation Techniques—A Review,Can. J. Expl. Geophys., 2, 55–
489 82.
490
491 Breiman, Leo (2001). "Random Forests". *Machine Learning* **45** (1): 5–
492 32. doi:10.1023/A:1010933404324.
493

Brely, N. (2010) “The International Monitoring System”, CTBTO Preparatory Commission, http://ctbtcourse.files.wordpress.com/2010/10/overview-and-technologies-of-ims.pdf.

Cleveland, K. M. and C. J. Ammon (2013). Precise relative earthquake location using surface waves, *J. Geophys. Res., 118*, 2893-2904, Doi: 10.1002/jgrb.50146

Gibbons, S. J., and F. Ringdal (2004). A waveform correlation procedure for detecting decoupled chemical explosions, NORSAR Scientific Report: Semiannual Technical Summary No. 2–2004, NORSAR, Kjeller, Norway, 41–50.

Gibbons, S. J., and F. Ringdal (2005). The detection of rockbursts at the Barentsburg coal mine, Spitsbergen, using waveform correlation on SPITS array data, NORSAR Scientific Report: Semiannual Technical Summary No. 1–2005, NORSAR, Kjeller, Norway, 35–48.

Gibbons, S., and F. Ringdal (2006). The detection of low magnitude seismic events using array-based waveform correlation, Geophys. J. Int. 165, 149–166.

Guttman, A. (1984). "R-Trees: A Dynamic Index Structure for Spatial Searching". *Proceedings of the 1984 ACM SIGMOD international conference on Management of data - SIGMOD '84*. p. 47.doi:10.1145/602259.602266.

Harris, D. B., 1991. A waveform correlation method for identifying quarry explosions, Bull. Seismol. Soc. Am. 80, no. 6, 2177-2193.

Harris, D. (2006), Subspace detectors: Theory, Lawrence Livermore Natl. Lab. Rep. UCRL-TR-222758, 46 pp., Lawrence Livermore Natl. Lab., Livermore, Calif.

Harris, D.B., and T. Kvaerna, 2010, Superresolution with seismic arrays using empirical matched field processing: Geophysical Journal International, v. 182, p. 1455-1477.

Harris, D. and D. Dodge (2011). An autonomous system for grouping events in a developing aftershock sequence, Bull. Seism. Soc. Am. 101, 763-774, doi:10.1785/0120100103.

Israelsson, H., 1990. Correlation of waveforms from closely spaced regional events, *Bull. Seism. soc. Am.*, **80**(6), 2177–2193.

Junek, W. N., VanDeMark, T. F., Saults, T. R., Harris, D. B., Dodge, D. A., Matlagh, S., Ichinose, G. A., Poffenberger, A., and R. C. Kemerait, 2013. “Integration of Empirical Signal Detectors into the Detection and Feature Extraction Application at the United States National Data Center”, poster, CTBTO Science and Technology Meeting, Vienna, Austria, 17-21 June, 2013.

Miners, Z. (2013). "Facebook's big data plans include warehouses, faster analytics", Computerworld, April 30, 2013.

Schaff, D. P., and P. G. Richards (2004). Repeating seismic events in China, Science 303, 1176-1178.

Schaff, D. P., and P. G. Richards (2011). On finding and using repeating seismic events in and near China. J. Geophys. Res. 116, doi:10.1029/2010JB007895.

Slinkard, M. E., Carr, D. B., and C. J. Young, 2013. Applying Waveform Correlation to Three Aftershock Sequences, *Bull. Seism. Soc. Am.*, 103(2A) 675-693; doi:10.1785/0120120058.

Tarvainen, M., and E. S. Husebye (1993). Spatial and Temporal Patterns of the Fennoscandian Seismicity – an Exercise in Explosion Monitoring, Geophysica, 29 (1-2) 1-19.

Van Trees, H. L. (1968), *Detection, Estimation and Modulation Theory*, vol. 1, John Wiley and Sons, New York.

Figure Captions

Figure 1 (A) shows the waveform completeness (number of events with waveforms per cell divided by the total number of events in the cell during the bounding epoch of the waveforms). Color is proportional to completeness with black lowest and white highest. Note that although the data set has global coverage, the completeness is highest in the Middle East, Eurasia, Fennoscandia, and Western North America. Panel (B) shows waveform segment counts by year and panel (C) shows the segment counts by year for waveform segments that eventually were found to correlate with another.

Figure 2 shows (schematically) the processing of an "island". The traversal strategy minimizes I/O and computations by requiring each waveform to be read only once and correlated only once with neighbors within 50 km. At each stage an R-tree is used to rapidly determine candidates. At the start, events 2-5 have been found to be within 50 km of (1). Waveforms for all five are loaded and (1) is processed against the others for all phases and bands. At this point, all data for (1) is removed from memory and the focus shifts to (2). Processing of the island continues until all events have been processed.

Figure 3 is a schematic illustration of the processing applied to a single channel for a pair of events observed by a single station. The graphic in the upper left shows the geometry of the station and events to be processed. The graphics labeled "Band 1" and "Band 2" show the seismogram pair filtered into two different bands, and indicate (schematically) the windows for which correlations will be computed. For each window pair, the cross correlation function is computed and the max and its associated shift are recorded in the database. This is indicated schematically in the lower part of the figure.

Figure 4 shows examples of common artifacts that correlate well and that were removed in a post-processing step. (a1) is an apparent calibration pulse. (a2) is a comb function due to some kind of electrical malfunction. (a3) is an unidentified artifact (perhaps sensor tilting?) that is surprisingly common on some STA-CHAN. (a4) is a step function probably due to an electrical malfunction. (b) is an example of an artifact caused by filtering a signal into a narrow band that contains noise and with the intended signal well outside the band. The top shows the raw traces with a high frequency seismogram riding on low frequency noise. After filtering into the band containing the noise, the intended signal is gone and only the narrow band noise is left. The filtered signal will correlate quite well, but the result has no seismological significance.

Figure 5 shows the overall distribution of correlations by phase (A) and by frequency band (B). In (A) the labels on each "stick" indicate the phase and the average window length. For all windows except "Whole" the length was predetermined but subject to the constraint that the correlation window could not

run into the next phase. The length of the "Whole" window was determined based on the source-receiver distance. Although this window could be as long as 2000s, because most of the retained correlations are for relatively short distances, the average length for this phase is only 82s. Part (B) shows the number of retained correlations as a function of filter band. The vast majority are in short-period bands which is not too surprising since most of the correlations are for relatively short distances.

Figure 6 shows the correlation counts as a function of event-station separation for long period bands (A), mid period bands (B) and short period bands (C). At mid to long periods the dominant feature in the plots is a drop of about 3 orders of magnitude for distances greater than 8 to 10 degrees. From that point out to about 90 degrees the number of correlations stays relatively constant except for a bump between 35 and 51 degrees. Part (D) shows the geometry of several arrays whose observations produce the "bump" in correlation counts between about 35 to 51 degrees.

Figure 7 shows the magnitude differences (left) and the distribution of time separations (right) for correlated event pairs in our results. The data are divided into four bins based on the average magnitude of each event pair. Panel (A) shows results for average Mw <= 2. Panel (B) shows results for 2 < Mw (avg) <= 4. Panel(C) shows results for 4 < Mw (avg) <= 6, and panel (D) shows results for 6 < Mw (avg) <= 8.

Figure 8 is a comparison of STA-CHAN waveform commonality on a year-by-year basis. Panel (A) uses 2010 as the reference year. It was produced by computing the intersection of the sets of waveform STA-CHAN each year with the set of waveform STA-CHAN in 2010. Note that until 1990 there are only tens of channels in common, but the number rises quite rapidly after 1990. Panel (B) was produced using 1977 as the reference year. It is scaled the same as (A) to show the relative size of the two data sets. Panel (B) also shows that only a few tens of STA-CHAN are common between the two data sets.

Figure 9 shows the geographic distribution of correlated events color-coded according to correlation fraction. The correlation fraction is defined as the number of events in a cell that correlate with at least one other event divided by the total number of bulletin events in the cell for the time period in which there are waveforms in the cell. Panel (A) shows the correlation fraction for all events. The dashed white line outlines the largest region in which our waveform holdings are at least 80% complete. Panel (B) shows the correlation fraction computed using only events >= Mw 5. Panel (C) shows the fraction of events for which we have waveforms for stations within five degrees of the epicenters.

Figure 10 shows the fraction of correlated seismicity as a function of source-station separation in different magnitude ranges. Panel (A) shows the fraction of Catalog Events with Correlations in All Bands for 6 Mw Ranges. Panel (B) shows the fraction

of Catalog Events with Correlations in Short-period Bands for 6 Mw Ranges. Panel (C) shows the fraction of Catalog Events with High Correlations (C >= 0.8) in Short-period Bands for 6 Mw Ranges.

Figure 11 shows 80s-long seismograms recorded at KK01 for 15 events found to be mutually correlated in the 1-2 Hz band at the >= 0.6 level (average correlation was 0.75). The source-receiver separation was between 48 and 50 degrees, and the average correlation window length was ~35s.

Figure 12 is a map of seismicity from the LLNL combined bulletin color coded according to distance from the nearest IMS station or array. Colors range from black for distances greater than 50 degrees to white for distance = 0. The small-circles are of 12 degree radius and are centered on IMS stations or arrays. Based on previous results, this is the effective bounding distance at which a substantial fraction of correlated waveforms may be observed in high frequency bands. Panel (A) shows the IMS primary stations and panel (B) shows the results for all IMS stations and arrays.

Table Captions

Table 1 shows the phases for which correlations could be computed. In order for the phase to be used at a specific event-station, the event had to fall within the depth range specified by (MIN DEPTH, MAX DEPTH) and the distance to the station had to be within (MIN DELTA, MAX DELTA). The window starting positions were calculated using AK135 and extended from PRE-WIN SECONDS before the predicted arrival for NOMINAL WIN LENGTH seconds. In a case where a window would extend into another predicted phase, the window was truncated at the predicted onset of the following phase. For the phase 'Whole' the nominal window length was calculated as MIN(nominal, DELTA (km) / 3).

Table 2 shows the frequency bands for which correlations might be computed. The bands were chosen so that for any phase and distance there would be at least one band optimum for the signal. For each window pair to be processed only those bands supported by the seismogram sample rate and containing a minimum of 10 cycles at the band center were used.

Figures

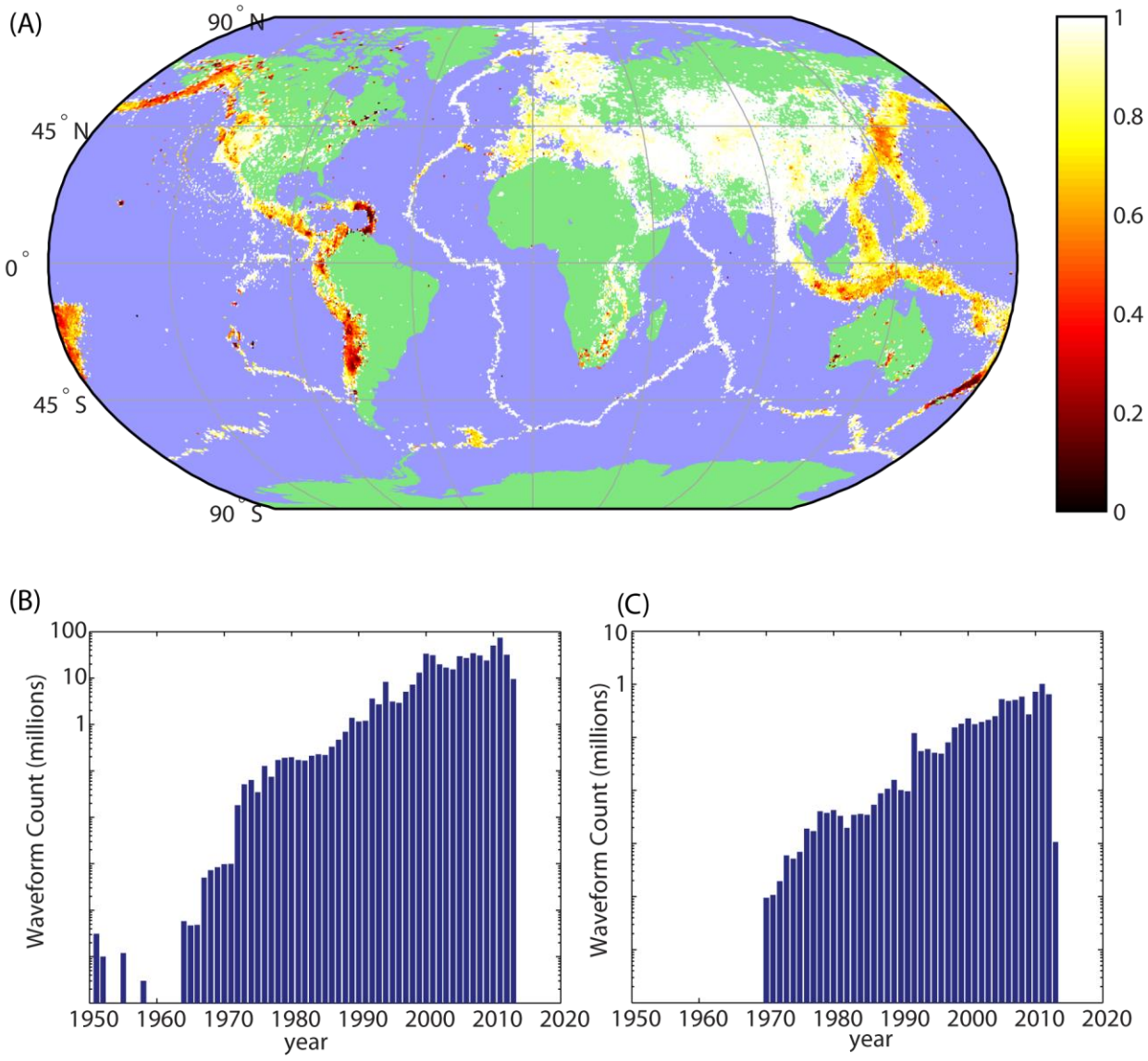
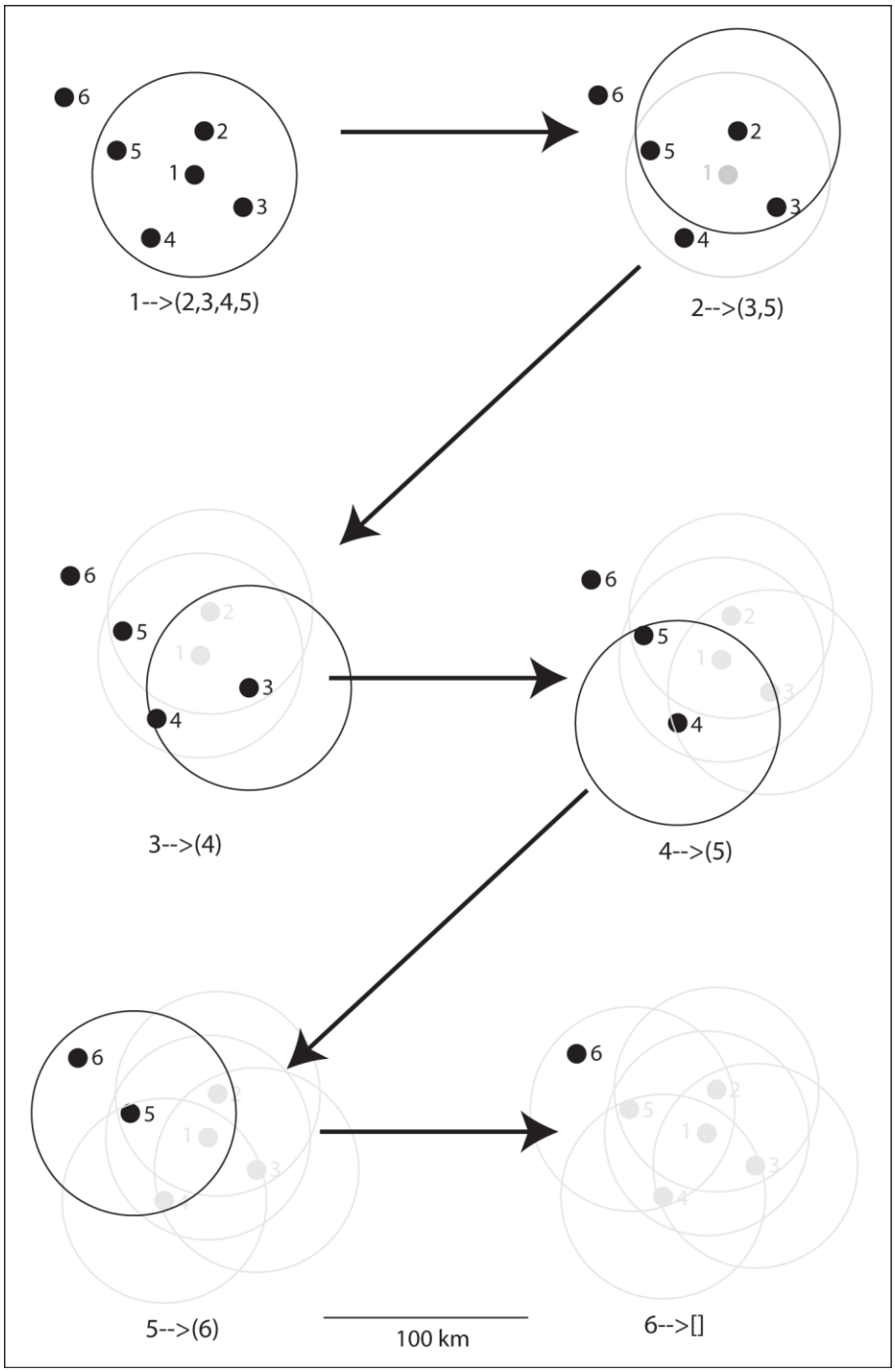
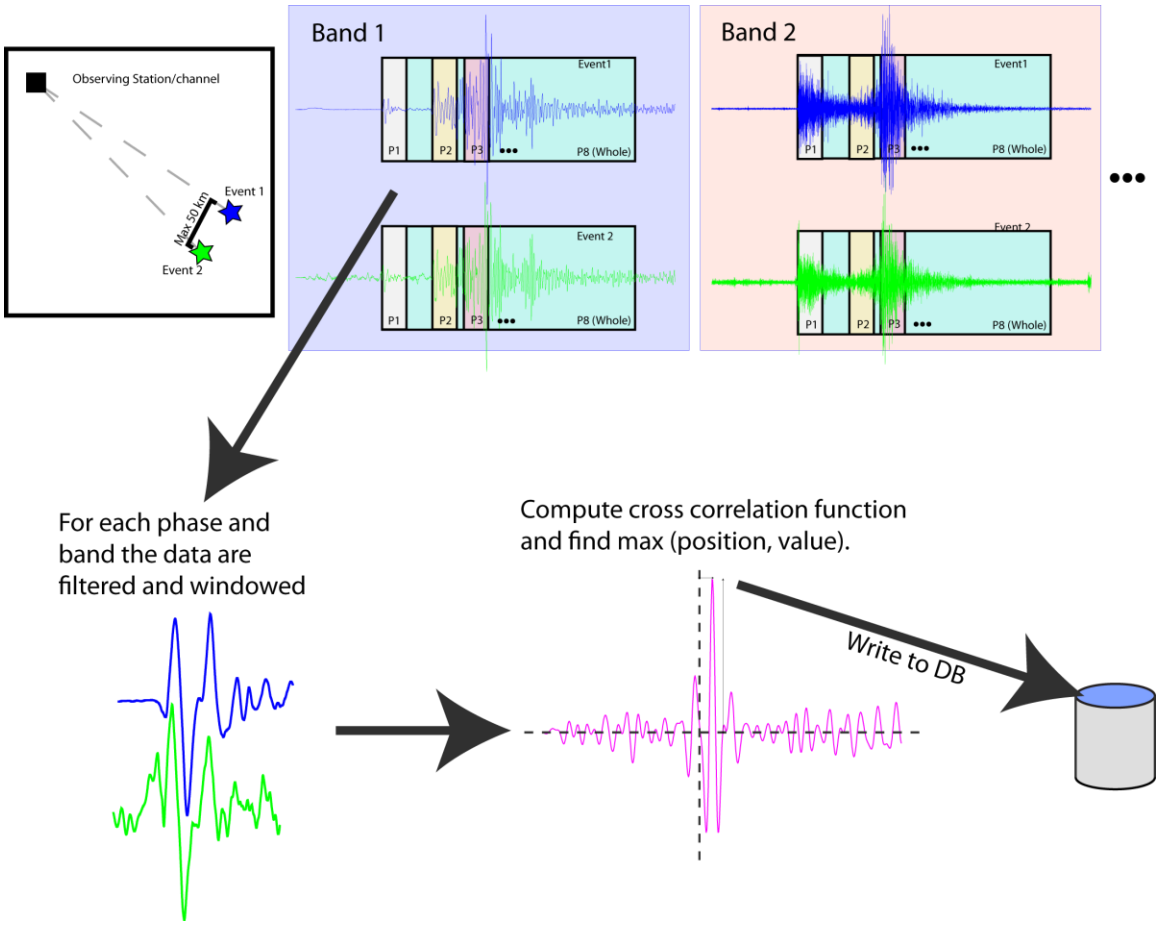


Figure 1.



688

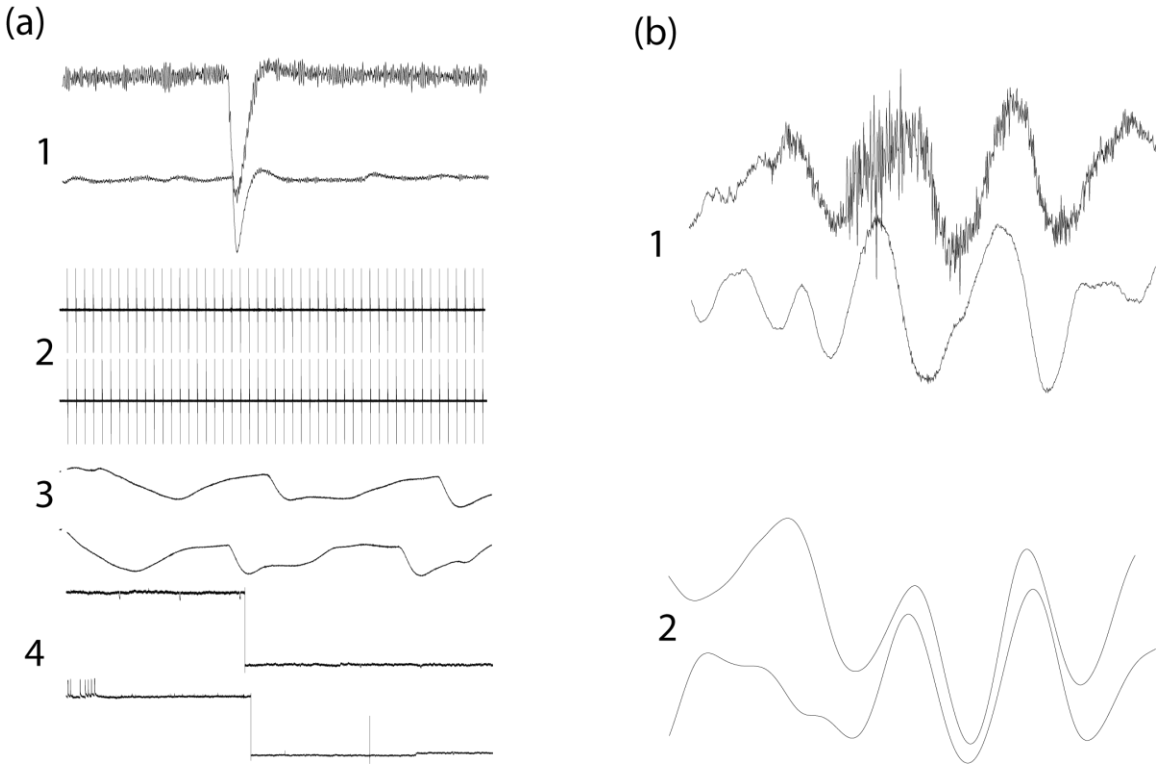
689 Figure 2.



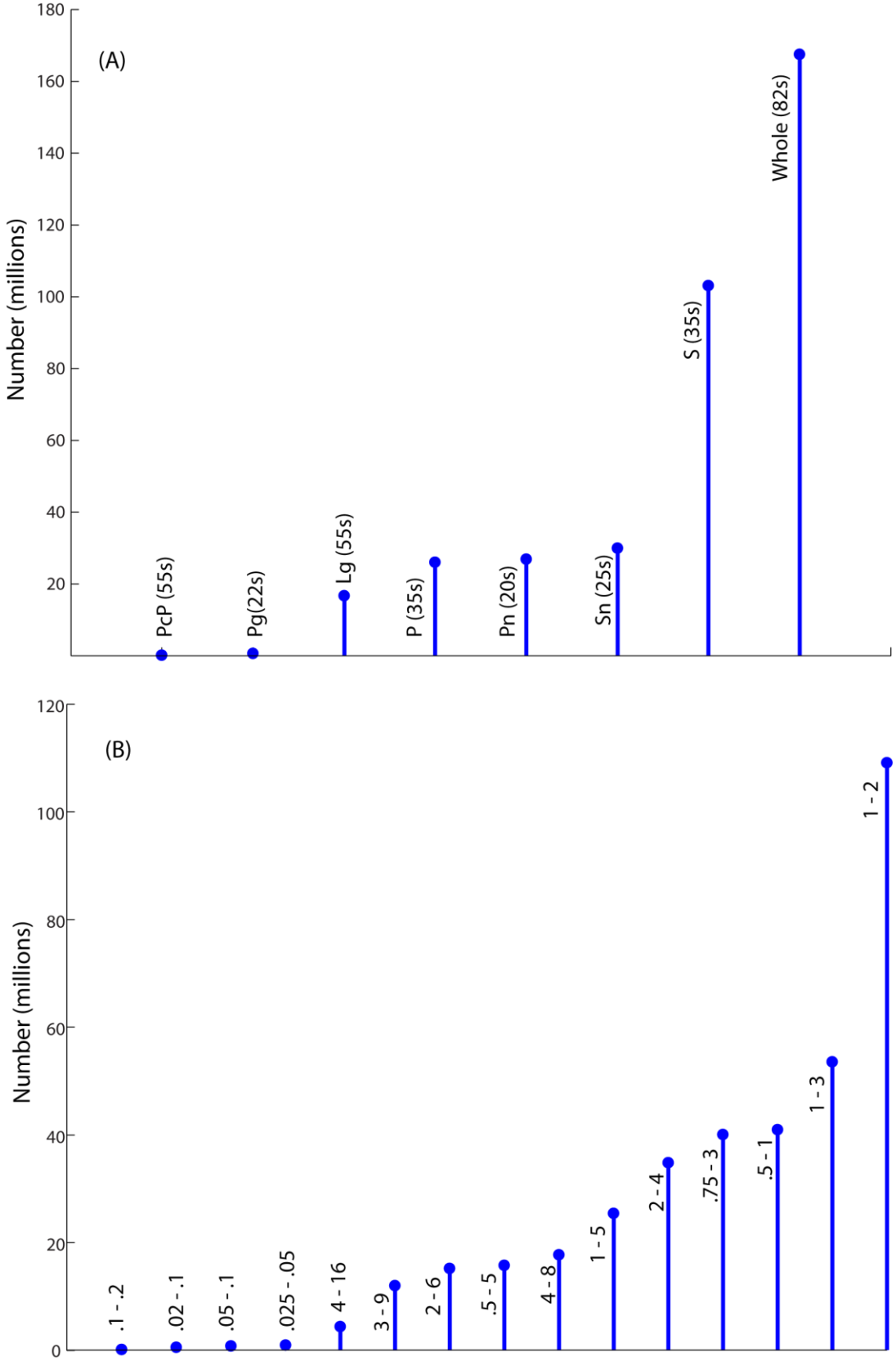
690

691 Figure 3

692



693
694 Figure 4.



695
696 Figure 5.

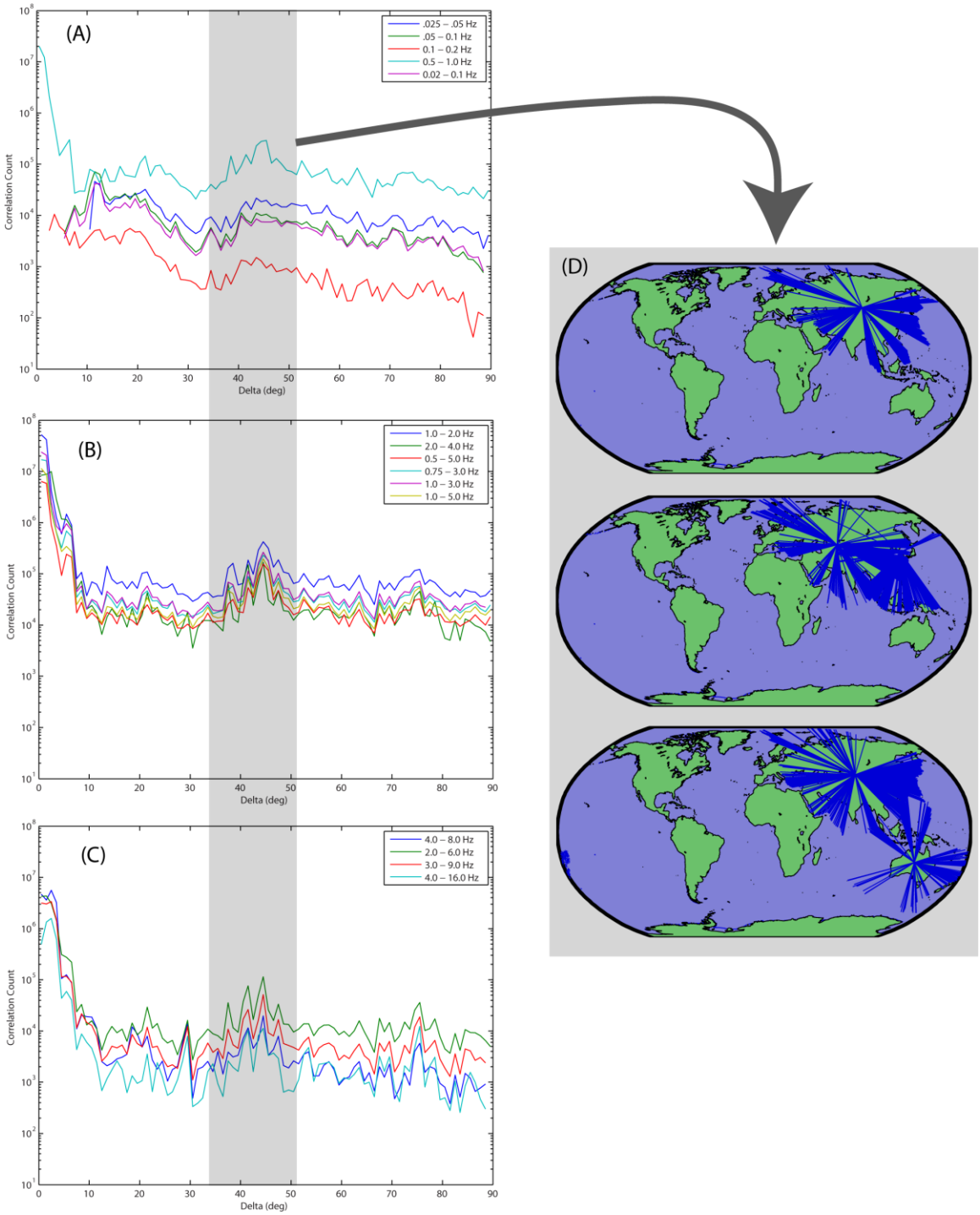
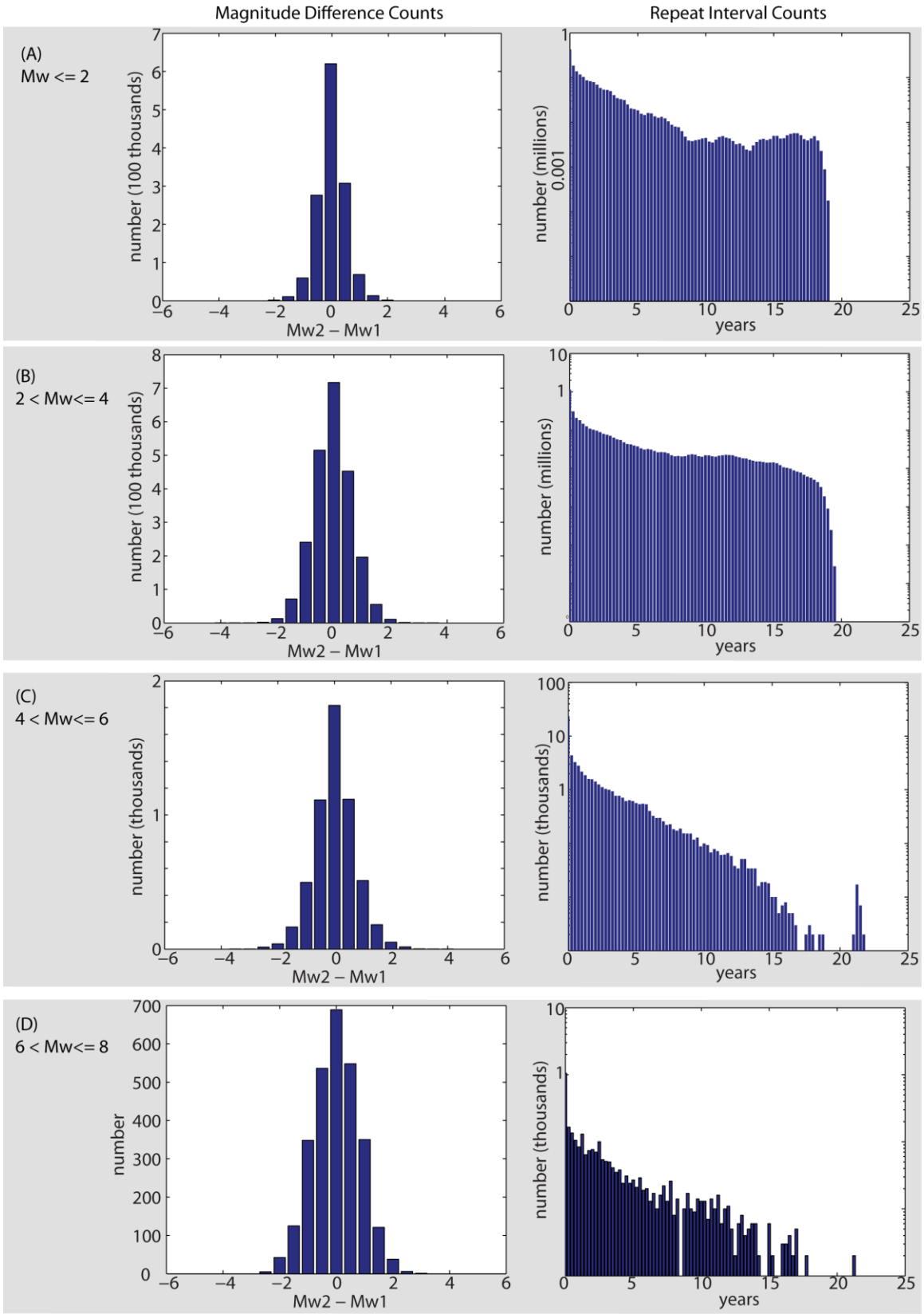
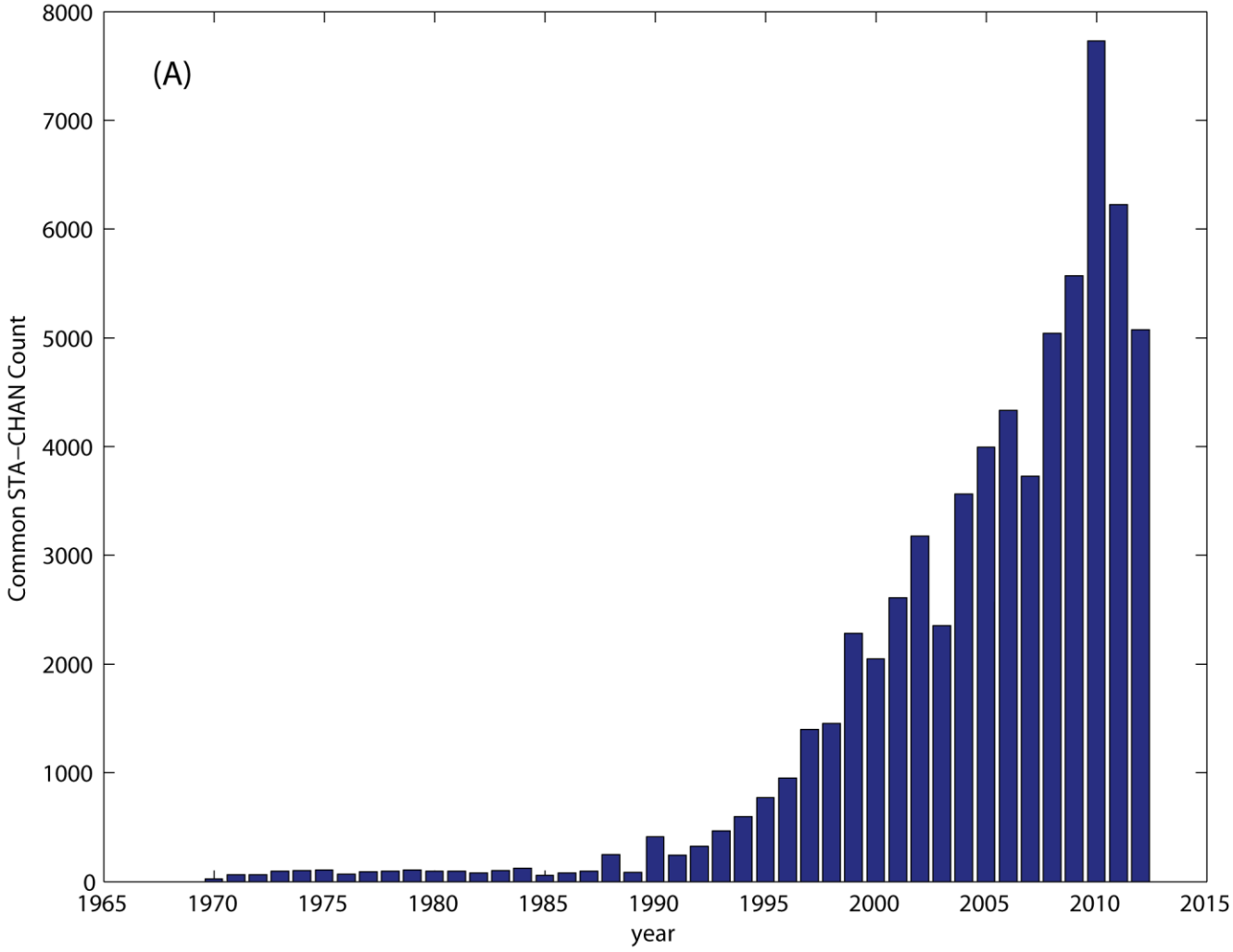
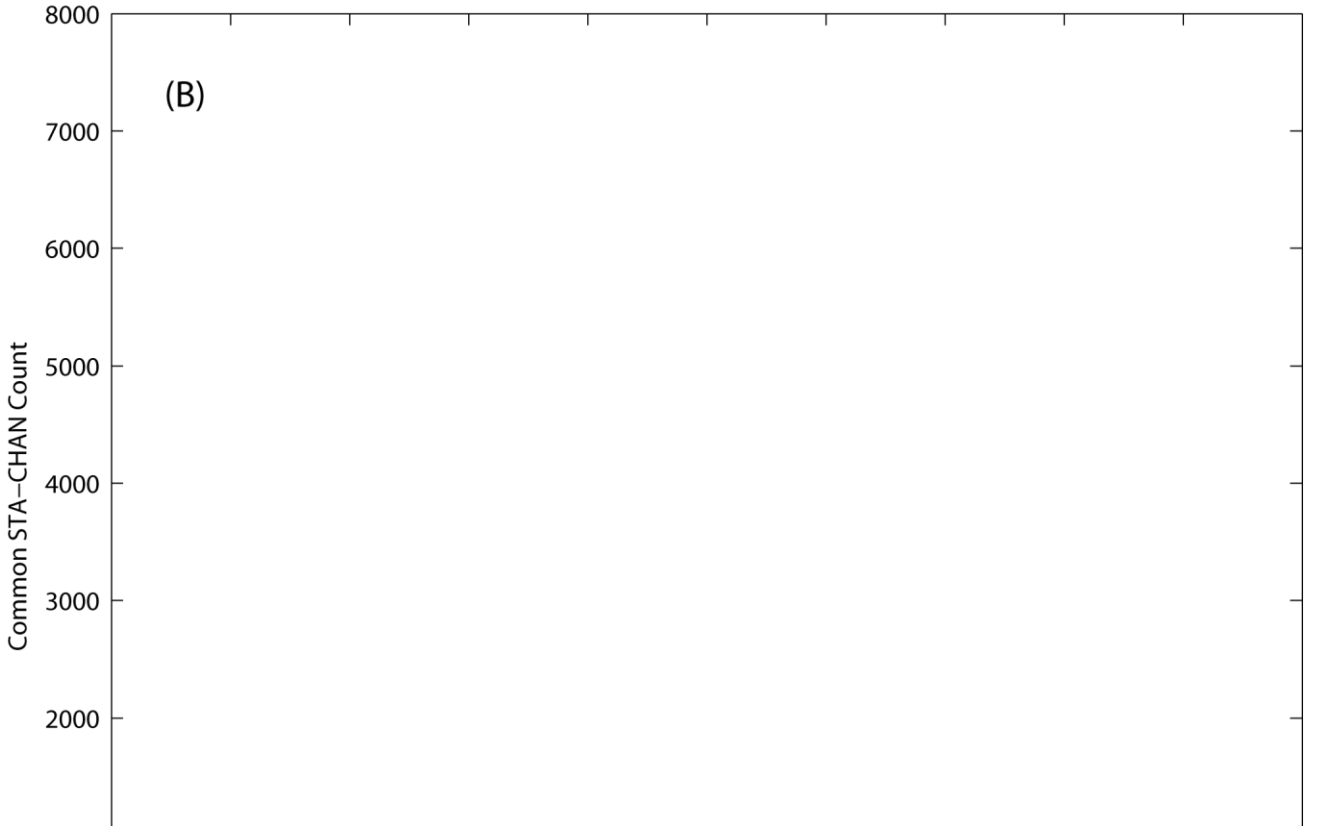

(A)
.025 – .05 Hz
.05 – 0.1 Hz
0.1 – 0.2 Hz
0.5 – 1.0 Hz
0.02 – 0.1 Hz
Correlation Count
Delta (deg)
(B)
1.0 – 2.0 Hz
2.0 – 4.0 Hz
0.5 – 5.0 Hz
0.75 – 3.0 Hz
1.0 – 3.0 Hz
1.0 – 5.0 Hz
Correlation Count
Delta (deg)
(C)
4.0 – 8.0 Hz
2.0 – 6.0 Hz
3.0 – 9.0 Hz
4.0 – 16.0 Hz
Correlation Count
Delta (deg)
(D)

Figure 6

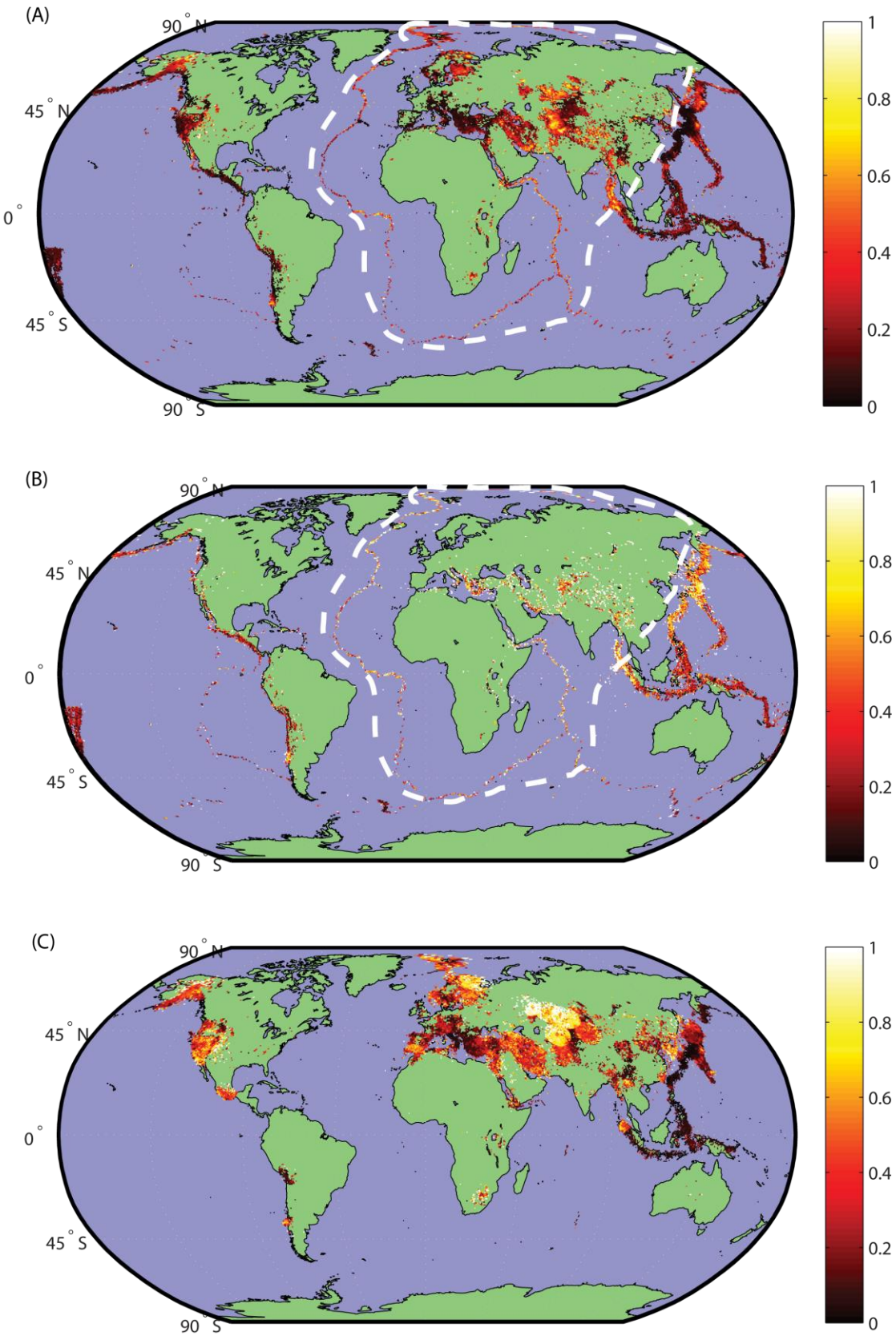


699
700 Figure 7
701


(A)
Common STA–CHAN Count
year
0
1000
2000
3000
4000
5000
6000
7000
8000
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015


(B)
Common STA–CHAN Count
2000
3000
4000
5000
6000
7000
8000

703 Figure 8
704



705
706 Figure 9

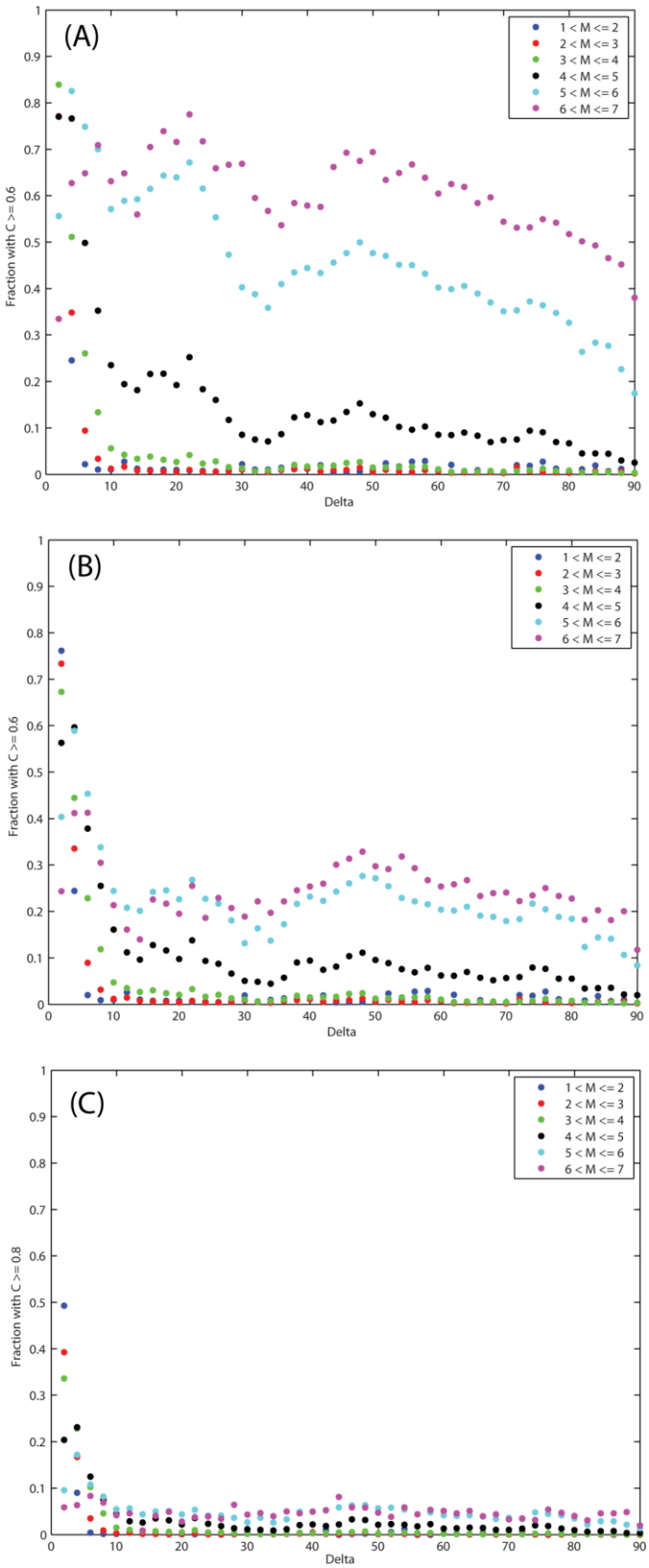

(A)
1 < M <= 2
2 < M <= 3
3 < M <= 4
4 < M <= 5
5 < M <= 6
6 < M <= 7
Fraction with C >= 0.6
Delta
(B)
1 < M <= 2
2 < M <= 3
3 < M <= 4
4 < M <= 5
5 < M <= 6
6 < M <= 7
Fraction with C >= 0.6
Delta
(C)
1 < M <= 2
2 < M <= 3
3 < M <= 4
4 < M <= 5
5 < M <= 6
6 < M <= 7
Fraction with C >= 0.8
Delta

Figure 10

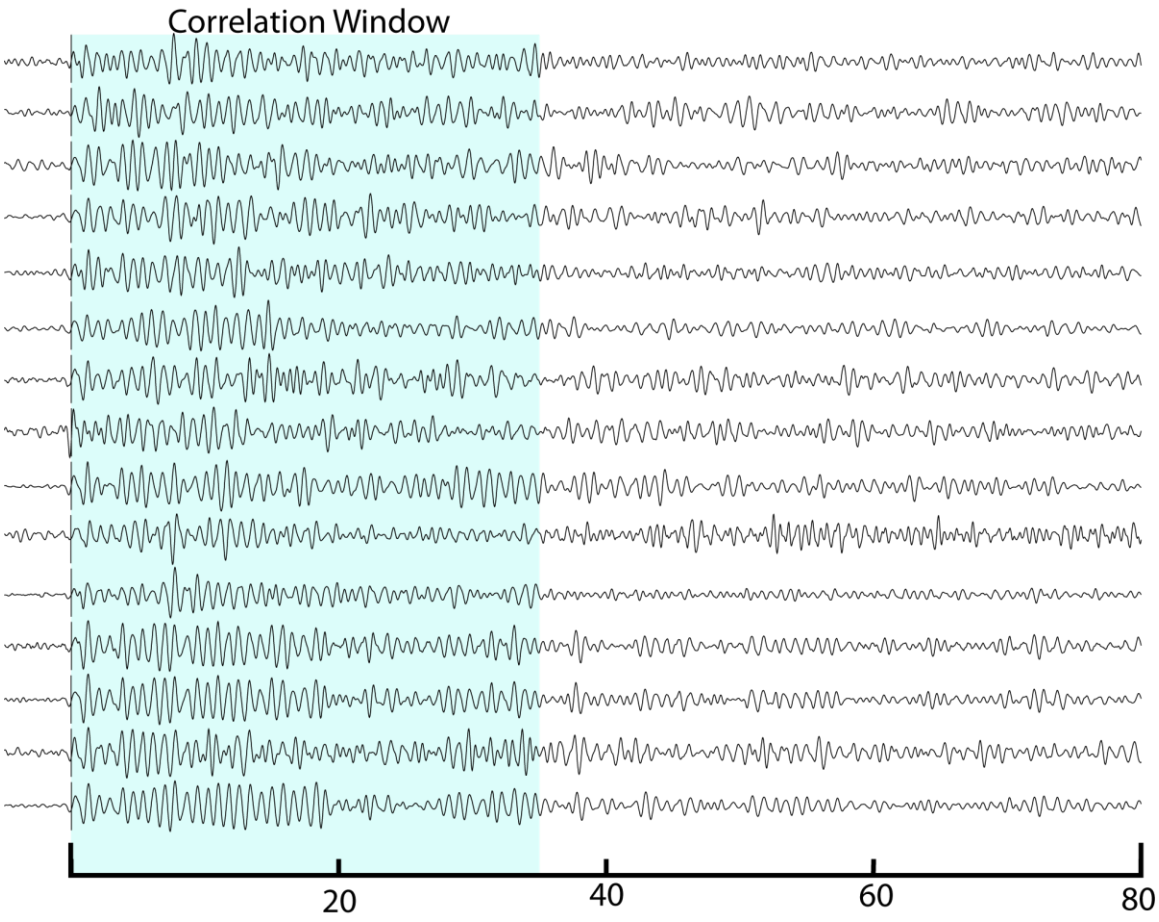
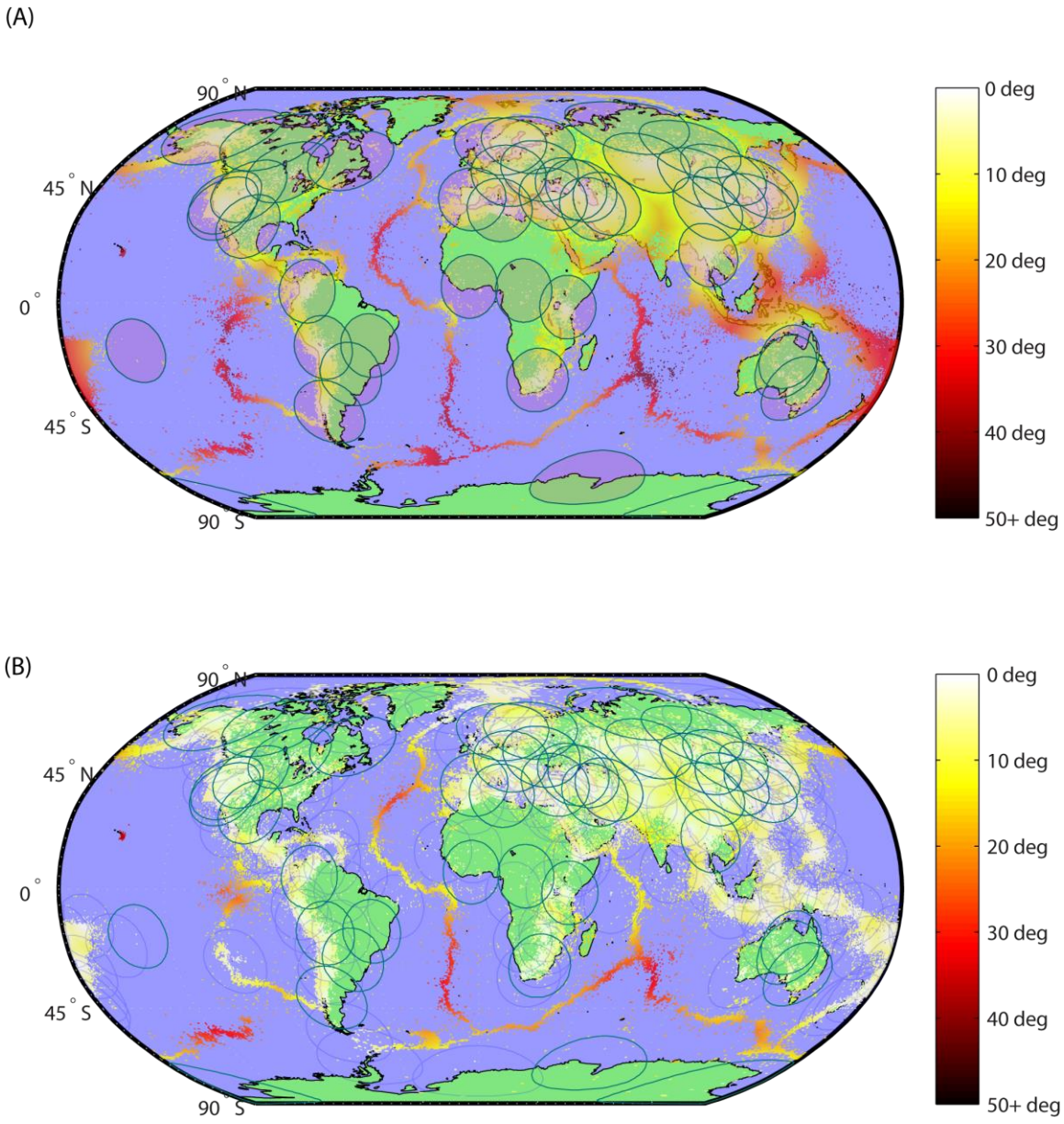


Figure 11



711
712
713 Figure 12
714
715
716

Tables

PHASE	NOMINAL WINDOW LENGTH (s)	PRE-WIN SECONDS	MIN Δ°	MAX Δ°	MAX DEPTH
Lg	50	10	1.46	15	35
P	30	5	0	90	700
PcP	50	5	26	60	700
Pg	30	10	0	1.5	35
Pn	15	7	1.5	10	35
S	30	10	0	90	700
Sn	30	10	1.46	15	35
Whole	2000	5	0	90	700

Table 1

LOW CORNER (Hz)	HIGH CORNER (Hz)
0.025	0.05
0.05	0.1
0.1	0.2
0.5	1
1	2
2	4
4	8
0.02	0.1
0.5	5
0.75	3
1	3
1	5
2	6
3	9
4	16

Table 2